Diffusion 기반 Text-to-Image모델의 이미지 역변환에 관한 연구

이수현, *박종열, *김혜지

서울과학기술대학교

hesyss@seoultech.ac.kr, *jongyoul@seoultech.ac.kr, *kimhyejee923@seoultech.ac.kr

A Study on the Image Inversion of Text-to-Image model based on Diffusion

Lee Soo Hyun, Park Jong Youl*, Kim Hye Ji*

Seoul National University of Science and Technology

요 약

Text-to-image 모델은 텍스트로부터 이미지를 생성하는 인공지능 모델이다. 해당 모델을 통해 가상의 이미지를 생성하는데 사용되는 요소들은 실제 이미지가 주어졌을 때 역변환을 통해서 찾게 된다. 해당 요소들은 Diffusion기반 Text-to-image 모델에서 이미지 편집을 위한 필수적인 정보이다. 기존의 역변환에 관한 연구들은 재구성된 이미지와 원본 이미지들에 대한 인공적인 결함이 발생하기도 하며, 기존의 파이프라인을 그대로 사용하지 못하는 경우가 있다. 본 논문에서 제안하는 기법은 기존의 사전학습된 생성형 인공지능 모델의 구조를 변함없이 사용할 수 있으며, 기존의 파이프라인을 있는 그대로 사용할 수 있기 때문에 더욱 범용적인 활용이 가능하다. 또한, 학습 파라미터를 낮춰 학습에 사용되는 시간을 단축하였다. 본 논문에서 제안하는 기법은 텍스트를 임베딩한 값 일부를 사용해 재구성 학습을 진행한다. 실험 부분에서 제안한 기법이 원본 이미지를 재구성하는데 기존의 기법과 비교하여 효율적인 기법임을 보였다.

Ⅰ. 서 론

최근 Stable Diffusion[1]이라는 Text-to-image(T2I) 모델에 관한 연구가 활발하게 이루어지면서 세간의 주목을 받고 있다. T2I는 사용자가 입력한 텍스트에 맞는 이미지를 인공지능 모델이 생성해주는 것을 말한다. 이제는 유튜브와 같은 소셜 플랫폼을 이용하여 누구나 본인의 방송을 만들고 쉽게 배포할 수 있다. 그러나 여전히 전문가를 거친 결과물과 그렇지 않은 결과물의 차이는 존재한다. 이에 따라 생성형 인공지능 중 T2I는 전문가의 손길을 거치지 않고 손쉽게 본인이 원하는 영상을 새로 만들어내거나 실제 존재하던 영상에 편집을 원하는 사용자의 요구에 발맞춰 발전해 나갈 것이다. 최근 사용되는 생성형 인공지능 기술들 또한 실제 이미지와 비교하여도 손색없는 고품질의 이미지를 만들어내는 것이 가능하고, 광고,일러스트, 삽화에 대한 가이드라인을 제공하기도 한다. 그러나 인공지능을 이용한 편집기술에는 여러 제한적인 부분이 있다. 생성된 이미지를 재가공할 때 발생하는 일관성 손실의 문제와 실제 이미지의 편집기술은 일련의 인공지능 모델을 통한 생성과정을 거치지 않았기 때문에 인공지능 모델을 이용한 편집을 위해서는 실제 이미지를 다시 재구성할 수 있는 정보들을 찾는 방법이 필요하다. 본 논문에서는 모델을 통해 생성한 이미지가 아닌 실제 이미지를 편집하기 위한 첫 단계인 이미지 재구성을 목적으로 한 기법을 소개한다.

Ⅱ. 관련연구

Stable Diffusion

T2I 모델 중 Stable Diffusion(SD)은 사용자가 입력하는 텍스트를 인코딩하는 CLIP-text encoder[2]를 사용하여 이미지 샘플 생성에 조건성을 부여하게 된다. 생성과정에서는 노이즈로부터 시작하여 조건화된 벡터(텍스트 임베딩)를 이용하여 텍스트에 맞는 이미지가 생성되기 위해서 제거되어야 할 노이즈를 예측하면서 추론이 진행되고 이러한 과정이 반복된다. 이를 활용한 편집기법들은 추론이 진행되면서 생성되는 값들에 변화를 주어 이미지의 객체를 바꾸거나, 배경에 변화를 주기도 한다. 반면에 실제 이미지는 사용자가 입력한 텍스트를 기반으로 생성된 이미지가 아니므로 이를 생성모델을 이용하여 편집하기 위해서는 이를 생성모델을 사용하여 재구성하여, 편집기법을 적용할 정보들을 얻는 것이 일반적이다. 수식(1)은 StableDiffusion에서 사용되는 추론식[3]이다.

$$\tilde{\epsilon}_\theta(z_t, t, C, \varnothing) = \omega \cdot \epsilon_\theta(z_t, t, C) + (1-\omega) \cdot \epsilon_\theta(z_t, t, \varnothing) \quad \dots(1)$$

$\tilde{\epsilon}_\theta$는 θ로 파라미터화 된 생성모델을 의미하며, z_t는 초기 노이즈 값, t는 해당 시점 C는 인코더를 거친 텍스트 임베딩, $\omega, \varnothing$는 각각 하이퍼파라미터, 빈 텍스트에 대한 임베딩 값을 의미한다. 빈 텍스트 부분은 비조건화 생성과정을 고려하고, 이를 이용해 텍스트에 등장하지 않은 내용을 제거하는데 영향을 준다.

Null-text Inversion

기존의 역변환 과정 중 하나인 Null-text Inversion[4]의 경우 $\varnothing$ 부분에 해당하는 임베딩들을 최적화 하여 이를 이용한다. 또한, 일반적인 방식으로는 Diffusion 모델의 특성상 단계적으로 발생하는 오류가 축적되기 때문에 이론적으로 최적화된 값을 알아도 추론과정으로 해당 지점에 도달 하는데 한계가 있음을 밝혔다. 따라서, 매 단계에서 발생하는 오류를 수정하기 위해 T단계의 추론과정이 있다면 T개의 미세조정된 임베딩을 사용하였다. 그러나 해당 방식은 기존의 빈 텍스트 부분이 이미지 생성에서 활용되던 원치 않는 이미지 생성을 방지하는 역할을 하지 못하게 된다. 본 논문은 이를 방지하기 위해 인코더를 거친 텍스트 임베딩을 분할하여 역변환하는 방법을 제안한다.

Ⅲ. 본 론

가. 배경지식

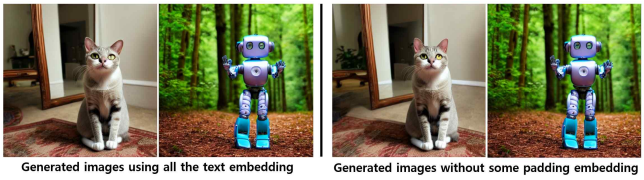

그림 1 모든 토큰을 사용해 생성한 이미지(좌)와 하위 2개의 토큰임베딩을 '0'으로 초기화 시킨 이미지(우)

본 논문에서는 텍스트 임베딩의 차원 중 일부를 사용하여 이미지의 재구성 기법을 제안한다. SD v1.4의 경우 텍스트 임베딩은 768차원으로 구성된 토큰이 총 77개 사용된다. 초기 텍스트는 Tokenizer를 거쳐 토큰화 되고 해당 토큰이 텍스트 인코더를 거쳐 768 차원으로 표현되게 된다. 만약 토큰 수가 제한을 넘어가면 생략되며, 제한 보다 적다면 패딩(문장의 끝을 알리는 토큰 'End Of Sentence')을 사용해 입력 차원을 유지하게 된다. 즉, 텍스트 임베딩은 의미를 갖는 임베딩과 패딩 임베딩으로 구분된다. 그림 1과 같이 패딩 임베딩은 토큰 후반부에 갈수록 생성된 이미지에 영향을 미치지 않는다. 이는 뒤로 갈수록 텍스트임베딩이 영상생성에 미치는 영향이 줄어든다는 것을 의미한다.

나. 제안 기법

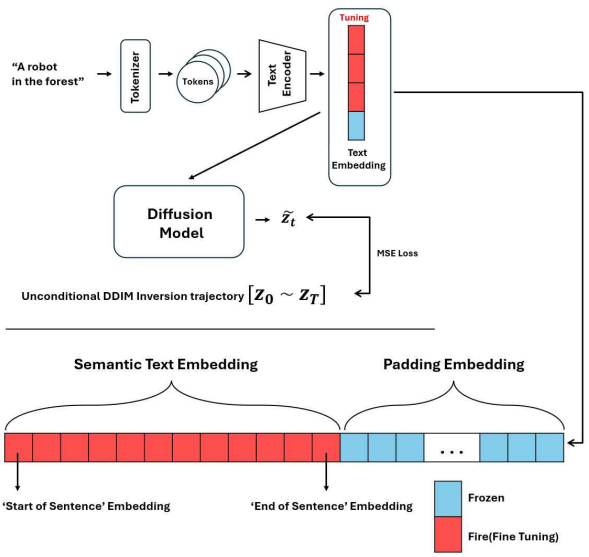

그림 2 전체적인 학습 프로세스(위)와 텍스트 임베딩의 내부 구성 및 학습 파라미터(아래)

Diffusion 모델은 단계적으로 노이즈를 제거하면서 이미지를 생성한다. 이때, 단계마다 노이즈가 제거된 값을 얻게 되는데 이를 잠재공간이라고 표현한다. 따라서, 총 T 단계를 거치며 이미지를 생성하면, T+1개의 잠재공간을 얻게 된다. Ⅲ-(가)에서 언급한바와 같이 영상생성은 텍스트 임베딩이 뒤로 갈수록 의미가 반영되지 않고, 인가된 텍스트가 영상생성에 대부분의 의미론적 정보를 가지고 있다. 따라서, 본 논문에서는 의미론적 텍스트 임베딩 부분의 값을 학습하여 영상 재구성하는 방법을 제안한다. 기존 방식[4]과 다르게 null-text '∅'가 아닌 텍스트 임베딩 부분을 최적화함으로 null-text 부분을 이용해 부정적(생성되지 않기를 바라는) 텍스트를 사용할 수 있다. 그림 2는 전체적인 제안 기법 흐름도를 보여준다. DDIM[5]에서는 비조건화의 이미지 생성에 대해서 역변환 과정은 생성과정을 역순으로 취하는 것과 동일하게 동작하며, 조건화 이미지가 아닌 경우에는 눈으로 확인하기 어려운 미세한 오차만 있기 때문에 이를 이용해서 구해진 값들을 정답 값으로 여긴다. 이 후 텍스트 임베딩의 의미론적 토큰 값만을 학습하면서 재구성 오차(MSE)를 줄이는 방향으로 학습이 진행된다.

Ⅳ. 실 험

가. 정성적 평가

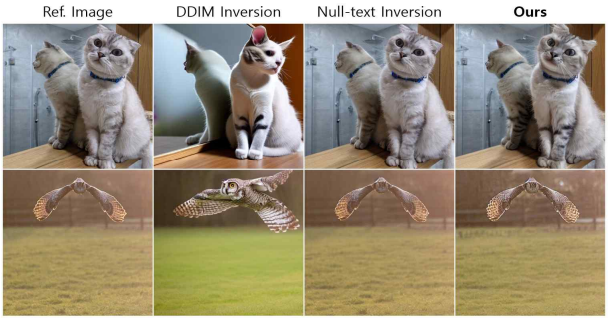

그림 3 이미지 재구성에 대한 정성적 평가. 좌측부터 원본이미지, DDIM Inversion, Null-text Inversion 그리고 우리가 제안한 기법이다.

그림 3의 정성적 평가를 보면 DDIM Inversion의 경우 원본이미지와 많이 다르나 Null-text Inversion과 우리가 제안한 기법은 원본과 매우 유사함을 보였다.

나. 정량적 평가

표 1은 제안한 기법과 기존 기법들에 대한 정량적 평가이다. #param은 갱신되는 파라미터의 수, step inv.는 재구성을 위한 반복횟수이며, CLIP-I의 경우, 원본이미지를 인코딩한 후 재구성된 이미지의 인코딩 벡터와의 유사도를 평가한다. LPIPS의 경우 두 이미지에 대한 인지적 거리를 측정하는 평가지표로 CLIP-I와 같이 원본이미지와의 유사도를 평가한다. 데이터 셋은 COCO2017 Validation[6]을 사용하였다.

	#param	step inv.	LPIPS(↓)	CLIP-I(↑)
DDIM Inv.	-	-	0.68	0.92
Null-text Inv.	2.95M	50	0.44	**0.98**
Ours	$l \times 50 \times 768$ **(Max=2.95M)**	**25**	**0.43**	0.97

표 1 이미지 재구성에 대한 정량적평가. l은 의미적 문장의 길이를 나타낸다.

LPIPS는 제안한 기법이 약간 우세하였으나, CLIP-I에서는 기존 기법이 좀 더 좋은 성능을 보였다. 그러나 매우 유사한 수준의 결과이며, 정성적이 평가로는 구분하기가 어렵다. 반면에 param, step을 낮춰 성능은 유지하되, 비용측면에서 우수함을 증명하였다.

Ⅴ. 결론 및 향후연구

본 논문에서는 이미지를 재구성하기 위한 역변환 기법을 제안하였다. 현재 주로 사용되는 [4]기법보다 더 적은 학습 파라미터와 단계를 거치면서 유사한 성능을 보임을 증명하였다. 역변환은 보다 좋은 성능의 이미지편집을 위한 첫 단계이므로 앞으로의 연구에서는 제안한 기법을 기반한 이미지편집에 대한 연구를 진행하며, 더 나아가서는 아직 명확한 이미지 평가지표가 없다는 점을 고려하여 통용될 수 있는 새로운 평가지표를 구축하기 위한 연구가 요구된다. 추가적으로 보다 다양한 데이터셋을 통한 검증이 요구된다. 해당 분야에 대한 연구는 디자인, 광고 등 다양한 미디어 종사자들에게 편의성을 향상시킬 가이드라인을 제공할 수 있는 기술이라고 여겨진다.

ACKNOWLEDGMENT

본 연구는 정부(교육부, 과학기술정보통신부)의 재원으로 일부 한국연구재단의 지원을 받아 수행(No.2024-0762 ,기여율: 50%)하고 일부 정보통신기획평가원의 지원을 받아 수행된 (IITP-2024-RS-2023-00262158, 기여율:50%) 연구임

참 고 문 헌

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[2] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PMLR, 2021.

[3] Ho, Jonathan, and Tim Salimans. "Classifier-free diffusion guidance." arXiv preprint arXiv:2207.12598 2022.

[4] M1okady, Ron, et al. "Null-text inversion for editing real images using guided diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[5] Song, Jiaming, Chenlin Meng, and Stefano Ermon. "Denoising diffusion implicit models." arXiv preprint arXiv:2010.02502 2020.

[6] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.